Janusz S. Bień
Katedra Lingwistyki Formalnej
Uniwersytet Warszawski

Dygitalizacja i komputeryzacja słowników na przykładzie *Słownika polszczyzny XVI wieku*

17 sierpnia 2009 r.,
15 lipca 2010 r., 23 sierpnia 2010 r.

Abstract

Digitalization and computerization of dictionaries (exemplified by the dictionary of the 16th century Polish)

For simple digitalization the recommended format is DjVu, which is produced in particular by the free `pdf2djvu` program. It is demonstrated that the OCR programs used by Kujawsko-Pomorska Digital Library are not adequate for the text of the dictionary, which contains many 16th century quotations. It is also stressed that scanning printed editions of digitally-born publications makes no sense from a user point of view, because the quality of OCRed scans will be always lower than that of an electronic document.

More sophisticated digitalization, called computerization, is only briefly mentioned, as the topic is discussed in details in Krzysztof Szafran's book. The possibility of treating the dictionary as a corpus is illustrated by searching the 32th volume of the dictionary with Poliqarp corpus tool.

1. Wstęp

Przedstawione tutaj prace nie byłyby możliwe bez życzliwości ś.p. Profesora Franciszka Pepłowskiego (1921 — 2009), wieloletniego kierownika Pracowni Słownika Polszczyzny XVI w. w Instytucie Badań Literackich PAN. Kontakty z Profesorem nawiązałem we wrześniu 2001 r., kiedy towarzyszyłem Zygmuntowi Saloniemu w jego wizycie w toruńskiej pracowni słownika. W wyniku dokonanych wówczas ustaleń uzyskałem (razem z kolegą Krzysztofem Szafranem) dostęp do plików komputerowych z tekstem ostatnich tomów słownika, a także do instrukcji redakcyjnej. Pliki komputerowe przekazywał nam jeden z redaktorów słownika Krzysztof Opaliński, który — wobec braku wsparcia ze strony zawodowego informatyka[1] — z zamiłowania i konieczności zajmował się również sprawami informatycznymi Pracowni. Bardzo pomocna była również kierownik pracowni wrocławskiej Małgorzata Nobis — w szczególności odnalazła ona pliki tomów, które omyłkowo nie zostały zachowane w pracowni toruńskiej.

W roku 2003 pracownia stanęła przed koniecznością zmiany systemu składania tekstów. Wybrany przez Opalińskiego system KOMBI (`http://www.3n.com.pl/kombi.php`) był wówczas klasycznym systemem WYSIWYG. Termin ten to skrót od *what you see is what you get* czyli *co widzisz, to dostaniesz*. Skrót ten bywa również rozwijany *what you see is all what you get* (*dostaniesz tylko to, co widzisz*), ponieważ systemy takie tworzą pliki komputerowe, które są przeznaczone wyłącznie do wydruku, i wykorzystanie ich do innych celów jest trudne. Przeciwko temu wyborowi zaprotestowałem stanowczo w piśmie do Prof. Pepłowskiego z 31 marca 2003 r. W wyniku intensywnej (prowadzonej pocztą elektroniczną) dyskusji z Opalińskim, do której włączył się autor systemu Stefan Nawrocki, możliwości systemu KOMBI zostały rozszerzone, o czym Nawrocki poinformował mnie 19 maja 2003 r. pisząc

[1] Dyrekcja IBL zapewniła Pracowni takie wsparcie dopiero w 2009 r., być może nie bez związku z moimi pismami w sprawie słownika kierowanymi między innymi do Fundacji na rzecz Nauki Polskiej jako jego sponsora; niektóre z nich są obecnie (23 sierpnia 2010 r.) dostępne publicznie ze strony wspomnianej dalej wyszukiwarki leksykograficznej.

http://bc.klf.uw.edu.pl

> Otóż zrobiłem konwerter do formatu znacznikowego. Właśnie przygotowuję się do aktualizacji programu do kolejnej wersji i mam już przygotowany plik, w którym ten format opisuję (w załączeniu).

Swoje zastrzeżenia do systemu KOMBI wycofałem już wcześniej, bezpośrednio po zapadnięciu decyzji o tym istotnym rozszerzeniu. W piśmie z 29 kwietnia 2003 r. (liczba dziennika 38/2003) Prof. Pepłowski pisał więc do mnie

> Cieszę się, że jednak uda się wykorzystać program KOMBI [...]
> Życzę sukcesów w zabieganiu o grant i w realizacji wielkiego planu internetowego.

Warto odnotować, że — według relacji Krzysztofa Opalińskiego — Stefan Nawrocki jako właściciel firmy 3N i twórca programu KOMBI wykazał daleko idącą życzliwość wobec potrzeb redaktorów słownika, bez czego wykonywany w redakcji sklad słownika byłby znacznie trudniejszy.

2. Dygitalizacja a komputeryzacja tekstów

Dygitalizacja (digitalizacja) to przekształcanie do postaci cyfrowej, w szczególności tekstów drukowanych lub pisanych na papierze. Wynik dygitalizacji może mieć mniej lub bardziej wyrafinowaną postać. Jednak od wielu lat coraz więcej tekstów, w szczególności publikacji, jest już w trakcie tworzenia wprowadzana do komputera, i postać drukowana jest dla nich wtórna. W przypadku takich „urodzonych cyfrowo" (ang. digitally born) tekstów lepiej mówić o ich komputeryzacji, to znaczy o przekształceniu do postaci bardziej wygodnej dla użytkowników. Kiedy mówi się o dygitalizacji dziedzictwa kulturowego, do którego należą w szczególności słowniki, należy mieć na myśli zarówno dygitalizację właściwą, jak i komputeryzację.

Formułowaniem rekomendacji dotyczących transformacji tekstów pisanych na elektroniczne — które w dużym stopniu są przydatne przy konwersji tekstów urodzonych cyfrowo — zajmuje się istniejące od 1987 r. konsorcjum *Text Encoding Initiative* (TEI), w swobodnym tłumaczeniu *Inicjatywa dygitalizacji tekstów* — występujące w oryginalnej nazwie słowo *encoding* jest użyte w bardzo ogólnym znaczeniu, które za Wikipedią można opisać jako zmianę formatu informacji, co trudno zwięźle oddać po polsku.

Rekomendacje TEI (`http://www.tei-c.org/release/doc/tei-p5-doc/en/html`) w punkcie 9.5 rozróżniają 3 sposoby reprezentacji słowników:

1. Perspektywa typograficzna — dwuwymiarowy obraz strony, z zachowaniem informacji o podziale na wiersze i strony.
2. Perspektywa edytorska — jednowymiarowy ciąg znaków, który może stanowić dane wejściowe dla systemu składania tekstów.
3. Perspektywa leksykalna — reprezentacja zawartej w słowniku informacji abstrahująca od jej formy tekstowej.

Jednocześnie można stosować więcej niż jedną taką perspektywę. Podział ten jest jednak bardzo zgrubny i nie wyczerpuje wszystkich możliwości.

Z inicjatywy amerykańskiej Federacji Bibliotek Cyfrowych w r. 1998 r. powołano grupy robocze analizujące przydatność rekomendacji TEI dla bibliotek cyfrowych. Jedna z nich (*TEI in Libraries Special Interest Group*, w skrócie TEILib) opracowała okresowo uaktualniany dokument *TEI Text Encoding in Libraries. Guidelines for Best Encoding Practices* (`http://www.diglib.org/standards/tei.htm`), który w wersji 2.1 z 2006 r. przewiduje 5 poziomów dygitalizacji:

1. Dygitalizacja całkowicie automatyczna (*Fully Automated Conversion and Encoding*).
2. Minimalna dygitalizacja edytowana (*Minimal Encoding*).
3. Dygitalizacja z prostą analizą dokumentu (*Simple Analysis*).
4. Dygitalizacja z podstawową analizą treści dokumentu (*Basic Content Analysis*)

http://rcin.org.pl

5. Cyfrowa edycja krytyczna (*Scholarly Encoding Projects*).

Rozróżnienia te można stosować również do tekstów, w których odpowiednia informacja jest zapisana w formaliźmie innym, niż rekomendowany przez TEI.

Najważniejszy jest dla nas tutaj poziom pierwszy, który może mieć różne formy, w szczególności mogą to być same obrazy wskanowanych stron, lub obrazy uzupełnione o niepoddany żadnej korekcie tekst będą wynikiem optycznego rozpoznawania znaków (*Optical Character Recognition*, OCR). Do poziomu pierwszego zaliczane są też odpowiednio skonwertowane teksty urodzone cyfrowe, co może być uzasadnione z bibliotecznego punktu widzenia, ale z punktu widzenia użytkownika — jak pokażemy dalej — teksty takie różnią się kolosalnie.

Piotr Żmigrodzki w swojej książce (2008:102–103) wyróżnia trzy stopnie dygitalizacji. Pierwszy stopień dygitalizacji to *plik graficzny możliwy do wyświetlania na ekranie komputera tak jak książka (bez jakiejkolwiek możliwości wyszukiwania w tekście*. Drugi stopień to pliki graficzne uzupełnione o pewne możliwości nawigacji. Trzeci stopień to oprogramowanie udostępniające tekst słownika w formie znakowej z możliwością wyrafinowego wyszukiwania i selekcji wyświetlanej informacji.

Każda z wymienionych klasyfikacji ma swoje słabe strony, w związku z tym większość słowników omawianych przeze mnie w artykułach (2006, 2009) trudno jest przypisać do jednej kategorii. W tej sytuacji pożyteczne jest także klasyfikowanie słowników z czysto technicznego punktu widzenia — dalej skoncentrujemy się na słownikach zapisanych w formacie DjVu.

3. Format i technologia DjVu

Jak pokazują statystyki polskiej Federacji Bibliotek Cyfrowych (`http://fbc.pionier.net.pl/owoc/attr-stats`), ponad 80% publikacji w 37 bibliotekach (stan na dzień 17 kwietnia 2009 r.) jest zapisanych w formacie DjVu. Są jednak również biblioteki, które nie stosują zupełnie tego formatu. Zachęcam do samodzielnego wyrobienia sobie opinii, czy biblioteki te wybrały właściwą drogę, przez porównanie dygitalizacji słownika Lindego (dostępny w formacie DjVu w Kujawsko-Pomorskiej Bibliotece Cyfrowej i w innym formacie w Polskiej Bibliotece Internetowej) i *Słownika Geograficznego Królestwa Polskiego i innych krajów słowiańskich* (dostępny w całości w formacie DjVu w Małopolskiej Bibliotece Cyfrowej i w innym formacie w Domenie Internetowych Repozytoriów Wiedzy, a pojedyncze hasło Wilno jest dostępne również w Cyfrowej Bibliotece Narodowej).

Choć format DjVu został specjalnie zaprojektowany do udostępnianie tekstów przez Internet, nie jest on bezpośrednio obsługiwany przez przeglądarki Internetowe i niezbędne jest zainstalowanie dodatkowego oprogramowania. Mówiąc ściślej, niektóre biblioteki cyfrowe oferują konwersję „na poczekaniu" formatu DjVu na format akceptowany przez przeglądarki, ale taka konwersja gubi zalety technologii DjVu — obejmuje ona oprócz samego formatu także wyrafinowane wykorzystanie pamięci podręcznej, pozwalające w sposób natychmiastowy wykonywać takie podstawowe operacje jak skalowanie (ang. *zoom*) czy pozycjonowanie (ang. *panning*).

Bardzo istotną cechą formatu jest to, że każda strona może być przechowywana w osobnym pliku i pobierana z serwera niezależnie od innych — jest to szczególnie przydatne w przypadku słowników, które nie są czytane w całości i po kolei.

Z naukowego punktu widzenia bardzo poważną zaletą DjVu jest możliwość swoistego cytowania fragmentów publikacji — poniższy adres internetowy wskazuje fragment słownika omawiany w dalszej części artykułu: `http://kpbc.umk.pl/Content/34034/Czytelnia_011_12.djvu?djvuopts=&page=Czytelnia_011_12_031_0001.djvu&zoom=455&showposition=0.73,0.08&highlight=1261,2986,1017,146`. Został on utworzony przez przez zaznaczenie tekstu i wybranie odpowiedniej pozycji w menu podręcznym przeglądarki `djview4`.

http://rcin.uw.edu.pl

4. *Słownik polszczyzny XVI wieku* w formacie DjVu

Tom I słownika (A - Bany) został z mojej inicjatywy udostępniony w Internecie 9 kwietnia 2006 r. — por. `http://bc.klf.uw.edu.pl/40/`. Jest to tzw. wersja *KiP* nazywana tak od nazwy witryny *Komputery i Polszczyzna*, na której tom ten aktualnie się znajduje. Tekst słownika w formacie DjVu został uzupełniony między innymi o spis treści w formie strony WWW w formacie HTML. Właściwy tekst słownika można więc zaklasyfikować jako dygitalizację pierwszego stopnia według Żmigrodzkiego, ale ze względu na spis treści cała edycja lokuje się pomiędzy pierwszym a drugim stopniem.

Wersja ta stanowiła inspirację dla Bożeny Bednarek-Michalskiej (kustosz dyplomowany w Bibliotece Głównej Uniwersytetu Mikołaja Kopernika), odpowiedzialnej za Kujawsko-Pomorską Bibliotekę Cyfrową, do podjęcia zakończonych powodzeniem starań o umieszczenie w Bibliotece wszystkich 32 opublikowanych tomów słownika[2] (`http://kpbc.umk.pl/publication/17781`); ostatnio został udostępniony także zeszyt próbny. Są to klasyczne dygitalizacje pierwszego poziomu w sensie TEILib. Tomy te są udostępniane w formacie DjVu z wykorzystaniem tzw. warstwy tekstu ukrytego zawierającej wyniki „brudnego OCR" czyli optycznego rozpoznawania znaków nie poddanego żadnej ręcznej korekcie. W czysto technicznym sensie można więc ich tekst przeszukiwać, a fragmenty zaznaczać i wklejać do tworzonego przez siebie tekstu, jednak liczba przekłamań jest tak duża, że trudno je uznać za dygitalizację drugiego stopnia w sensie Żmigrodzkiego. Dygitalizacje te lokują się więc pomiędzy pierwszym a drugim stopniem, choć z innych powodów, niż edycja KiP. Jednocześnie brak jakichkolwiek pomocy nawigacyjnych nie pozwala ich zaliczyć do drugiego poziomu w sensie TEILib, stanowią za to najbardziej typowy przykład pierwszego poziomu dygitalizacji.

W przedstawionych wyżej przypadkach obrazy graficzne stron zostały uzyskane przez skanowanie, a warstwa tekstowa została utworzona w sposób wtórny. Format DjVu może być stosowany jednak również do publikacji urodzonych cyfrowo, dla których to warstwa graficzna jest wtórnie utworzona na podstawie tekstu właściwego, który staje się tekstem ukrytym. Umożliwia to w szczególności nowatorski program `pdf2djvu` (Wilk 2008), dostępny bezpłatnie na licencji GNU General Public License w większości ważniejszych dystrybucji systemu Linux, a od pewnego czasu także dla MS Windows.

5. Wykorzystanie plików poligraficznych

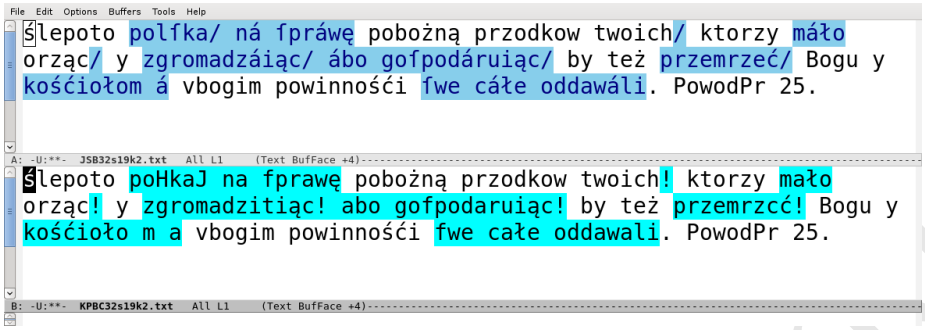
Przez pliki poligraficzne rozumiem pliki przekazywane do drukarni w celu sporządzenia matrycy lub bezpośredniego druku cyfrowego. Przez dłuższy czas podstawowym formatem tych plików był tzw. Postscript, obecnie jest to PDF (*Portable Document Format*), od pewnego czasu mający charakter oficjalnego międzynarodowego standardu ISO. W tym właśnie formacie zostały przekazane do drukarni tomy XXXI i XXXII słownika, ale ponieważ słownik jest drukowany z matryc, zawierały one lustrzane odbicie właściwego tekstu. Na moją prośbę Krzysztof Opaliński przygotował odpowiednie pliki XXXII tomu bez lustrzanego odbicia. Pliki te zostały następnie przetworzone za pomocą wspomnianego wyżej programu `pdf2djvu`, co na współczesnym komputerze PC wymaga podania jednej komendy i zaczekania około pół godziny na wynik. Niestety tak utworzona wersja 32 tomu w formacie DjVu, choć gotowa do udostępnienia od marca 2008 r., nie jest ciągle dostępna publicznie z powodu braku zgody dyrekcji Instytutu Badań Literackich (oficjalnie wystąpiłem o nią 6 maja 2008 r., na posiedzeniu w dniu 9 marca 2009 r. moją inicjatywę poparł Komitet Językoznawstwa PAN)[3].

[2] Aktualnie (23 sierpnia 2010 r.) są opublikowane i dostępne w KPBC 33 tomy.

[3] Dnia 14 lipca 2010 r. urodzone cyfrowo wersje tomów XXXI i XXXII zastąpiły w KPBC wersje wskanowane; pod względem typograficznym różnią się one nieznacznie od wersji drukowanych. Tom XXXIII też jest udostępniony w wersji urodzonej cyfrowo.

http://bc.klf.uw.edu.pl

Na ilustracji 1 widzimy porównanie fragmentów tekstu ukrytego w dwóch wersjach tomu XXXII — na górze wersja KiP bazująca na plikach poligraficznych, na dole wskanowana wersja KPBC.



Rysunek 1. Przykładowe błędy OCR w XXXII tomie słownika

Warto przede wszystkim zwrócić uwagę na błędy systematyczne — użyty program do OCR ma zbyt ubogi repertuar znaków. W konsekwencji długie *s* przechodzi zawsze na *f*, pomijane są też akcenty nad literami (w tym przykładzie konkretnie nad *a*), wprowadzana jest niepotrzebna spacja, kiedy część słowa jest zapisana we frakcji górnej (*kośćiołom*). Systematycznych błędem innego rodzaju jest rozpoznawanie ukośnej kreski jako wykrzyknika. Oprócz tego mylone są inne znaki o podobnych kształtach, stąd mamy np. *przemrzcć* zamiast *przemrzeć*. Sądząc po tym przykładzie, błędnie rozpoznana jest prawie połowa słów.

Reasumując, skanowanie urodzonych cyfrowo tekstów i wykonywanie dla nich OCR można określić krótko: gorzej! drożej! dłużej!

6. Wykorzystanie plików redakcyjnych

Pliki redakcyjne zawierają więcej informacji niż pliki poligraficzne i przez to stanowią lepszą podstawę do dygitalizacji wyższego stopnia. Zilustrujemy to tutaj tylko jednym przykładem: plik redakcyjny musi zawierać komendę uwzględnienia wyrazu hasłowego w żywej paginie, co pozwala łatwo zautomatyzować sporządzanie indeksu haseł. Rozpoznanie struktury hasła też jest łatwiejsze w plikach redakcyjnych niż w plikach poligraficznych. O eksperymencie polegającym na wykorzystaniu wspomnianego wcześniej formatu znacznikowego systemu KOMBI do stworzenia wersji elektronicznej tomu XXXI pisze Krzysztof Szafran w swojej książce (2007), która jest dostępna w elektronicznej bibliotece Uniwersytetu Warszawskiego — po rozwiązaniu umowy z wydawnictwem — na liberalnych zasadach licencji Creative Commons (wersja Uznanie autorstwa — Użycie niekomercyjne – Bez utworów zależnych 2.5 Polska).

Tworząc wersję elektroniczną na podstawie plików redakcyjnych możemy również przystosować format strony do proporcji ekranu i wprowadzić inne udogodnienia, na przykład wyróżniać kolorem poszczególne części artykułu hasłowego, co w druku jest nadmiernie kosztowne.

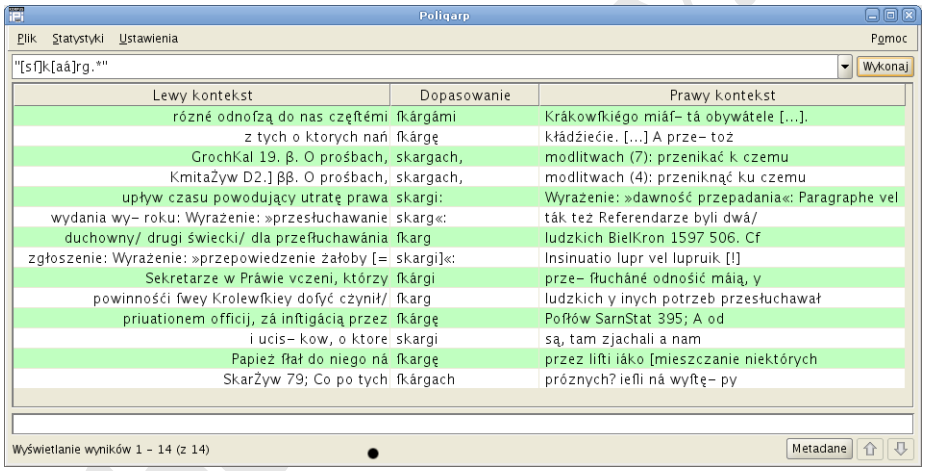
Konwersja plików redakcyjnych na edycję elektroniczną jest tym łatwiejsza, im więcej informacji zapisanych w nich jest w sposób jawny. Rozwiązaniem idealnym jest tworzenie plików redakcyjnych w formacie XML, a dobrym rozwiązaniem pośrednim jest wykorzystanie systemu TeX lub LaTeX. Przykładem tego drugiego podejścia jest przygotowany do druku za pomocą systemu TeX (w dużym stopniu przez autorów) słownik angielsko-polski Tadeusza Piotrowskiego i Zygmunta Saloniego, który doczekał się trzech różnych wersji elektronicznych w dużym stopniu dlatego, że ich tworzenie było względnie łatwe — por. np. (Głowińska, Woliński 2000).

http://rcin.kul.lublin.edu.pl

7. Słownik jako korpus

Niestety brak miejsca nie pozwala ustosunkować się do ciekawego i ważnego artykułu (Żmigrodzki 2005), w którym postulat traktowania słownika jako korpusu pojawił się chyba po raz pierwszy. Z punktu widzenia użytkownika słownik jest dostępny jako korpus wtedy, kiedy do jego przeszukiwania stosuje się takie same narzędzia, jak do korpusu.

Jednym z takich narzędzi jest Poliqarp (*Polyinterpretation Indexing Query and Retrieval Processor*, w swobodnym tłumaczeniu *Procesor kwerend i wyszukiwań z indeksowaniem wielointerpretacyjnym*), stosowany między innymi w Narodowym Korpusie Języka Polskiego (`http://nkjp.pl/`). W ramach moich zajęć *Reprezentacja tekstów w systemach komputerowych* student informatyki p. Piotr Sikora wykonał prototyp konwertera z formatu DjVu na format XCES (*XML Corpus Encoding Format*), co pozwala wykorzystać program Poliqarp do przeszukiwania tekstu słownika.[4]. Rozwiązanie to jest aktualnie testowane na XXXII tomie słownika Na ilustracji 2 widzimy przykład kwerendy wykorzystującej tzw. wyrażenia regularne. Pokazuje ona konkretnie, jak szukać słów zaczynających się od napisu *skarg* uwzględniając warianty pisowniowe: *s* może być zarówno długie, jak i krótkie, zaś *a* z kreską lub bez.



Rysunek 2. Przeszukiwanie słownika za pomocą programu Poliqarp

8. Podsumowanie

Choć przedstawione wyżej metody dygitalizacji są ilustrowane przykładami ze *Słownika polszczyzny XVI wieku*, mają one charakter ogólny i mogą być stosowane do różnych aktualnie opracowywanych słowników. Są to metody nie tylko łatwe, ale i tanie, ponieważ wszystkie niezbędne narzędzia są lub będą dostępne bezpłatnie na licencji GNU General Public License lub podobnej.

Prace cytowane

Bień, J.S., 2006. Kilka przykładów dygitalizacji słowników *Poradnik Językowy* z. 8 (637), s. 55-63. `http://ebuw.uw.edu.pl/publication/250`.

[4] Obecnie (23 sierpnia 2010 r.) do tego celu jest używana specjalna wersja programu Poliqarp — a dokładniej jego klienta —przygotowana przez Jakuba Wilka. Tzw. wyszukiwarka leksykograficzna jest dostępna pod adresem `http://poliqarp.wbl.klf.uw.edu.pl/` i obsługuje wszystkie dostępne tomy *Słownika polszczyzny XVI wieku* oraz słownik warszawski, a od niedawna także drugie wydanie słownika Lindego.

http://bc.ibl.waw.pl

—2009 Digitalizing dictionaries of Polish. [w:] *Methods of Lexical Analysis: Theoretical assumption and practical applications.* Białystok, s. 37-45. http://bc.klf.uw.edu.pl/71/

Głowińska, K., Woliński, M., 2000. Angielsko-polski słownik elektroniczny XeLDA. *Acta Universitatis Nicolai Copernici. Studia Slavica* 5, z. 343, s. 119-124. http://kpbc.umk.pl/Content/19304/AUNC_005_42.djvu?djvuopts=&page=AUNC_005_42_119.djvu

Szafran, K., 2007. *Analiza i formalny opis struktury „Słownika polszczyzny XVI wieku"*, Warszawa. http://ebuw.uw.edu.pl/publication/253

Wilk, J., 2008. Rozbudowa pakietu oprogramowania DjVuLibre. http://jw209508.hopto.org/papers/thesis/.

Żmigrodzki, P., 2005. Słownik jako korpus tekstów — korpus tekstów jako słownik. Perspektywy polskiej leksykografii naukowej. *Poradnik Językowy* nr 6, s. 3-14.

—2008. *Słowo — słownik — rzeczywistość*. Lexis: Kraków.

http://bc.klf.uw.edu.pl